Семинар по проектированию Dell Storage SC: гео-распределенные ЦОД и SSD в массивах

Федор Павлов

консультант по технологиям хранения

Гео-распределенные площадки

- Общий вид:
  - список вопросов и технических трудностей
- Концептуальная основа
  - DA vs DR
- Техника
- Лицензирование
- Материалы

Общий вид – вопросы и задачи

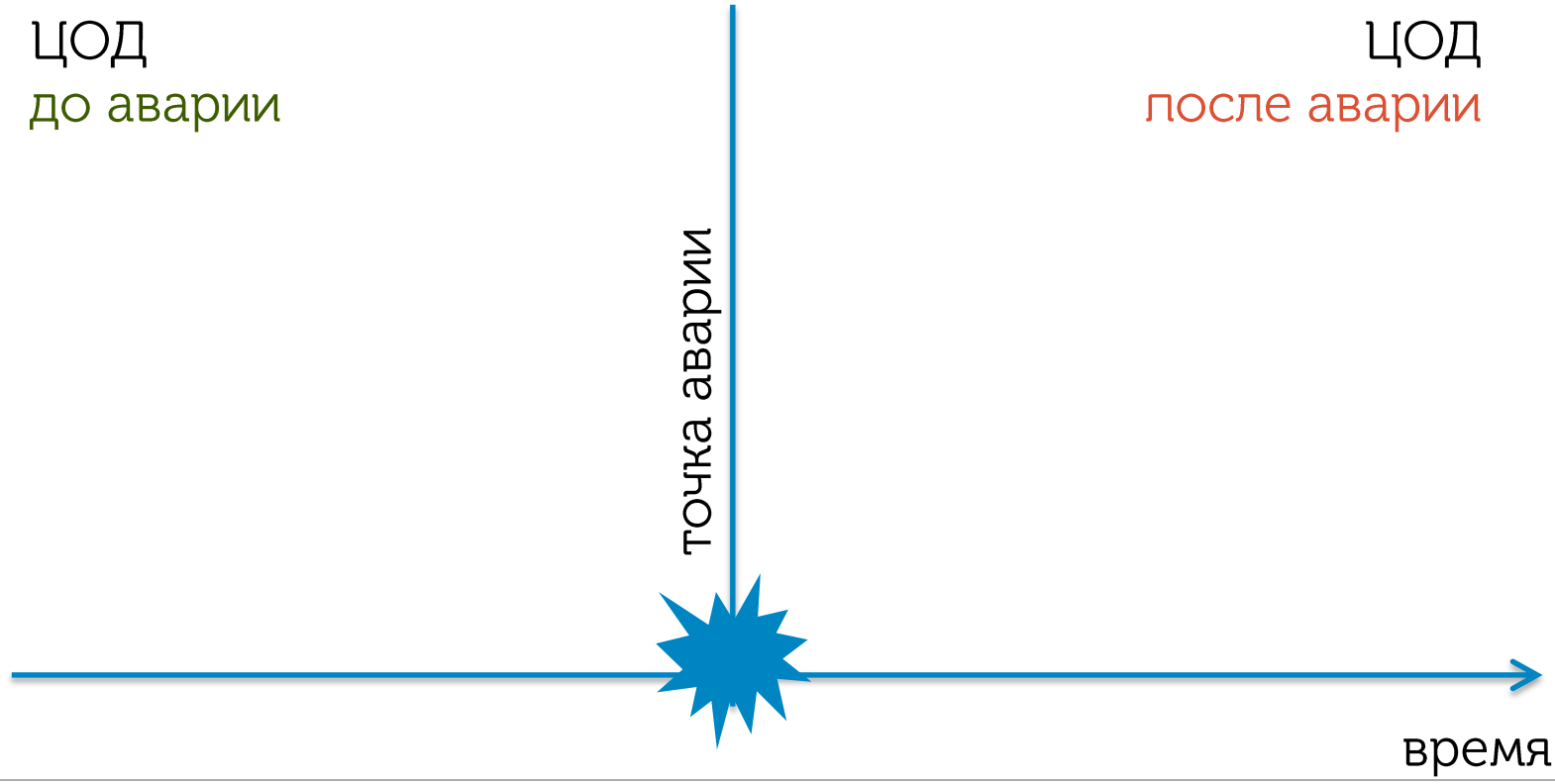
- Растянутый ЦОД
- Задержки в канале
- Split-Brain
- Способы переключения
- Поддержка ОС и гипервизоров
- Метрокластер и файловая система
- Стоимость
- Подготовка решения и внедрение
- Технические и экономические риски

Концептуальная основа:
Disaster Avoidance vs Disaster Recovery

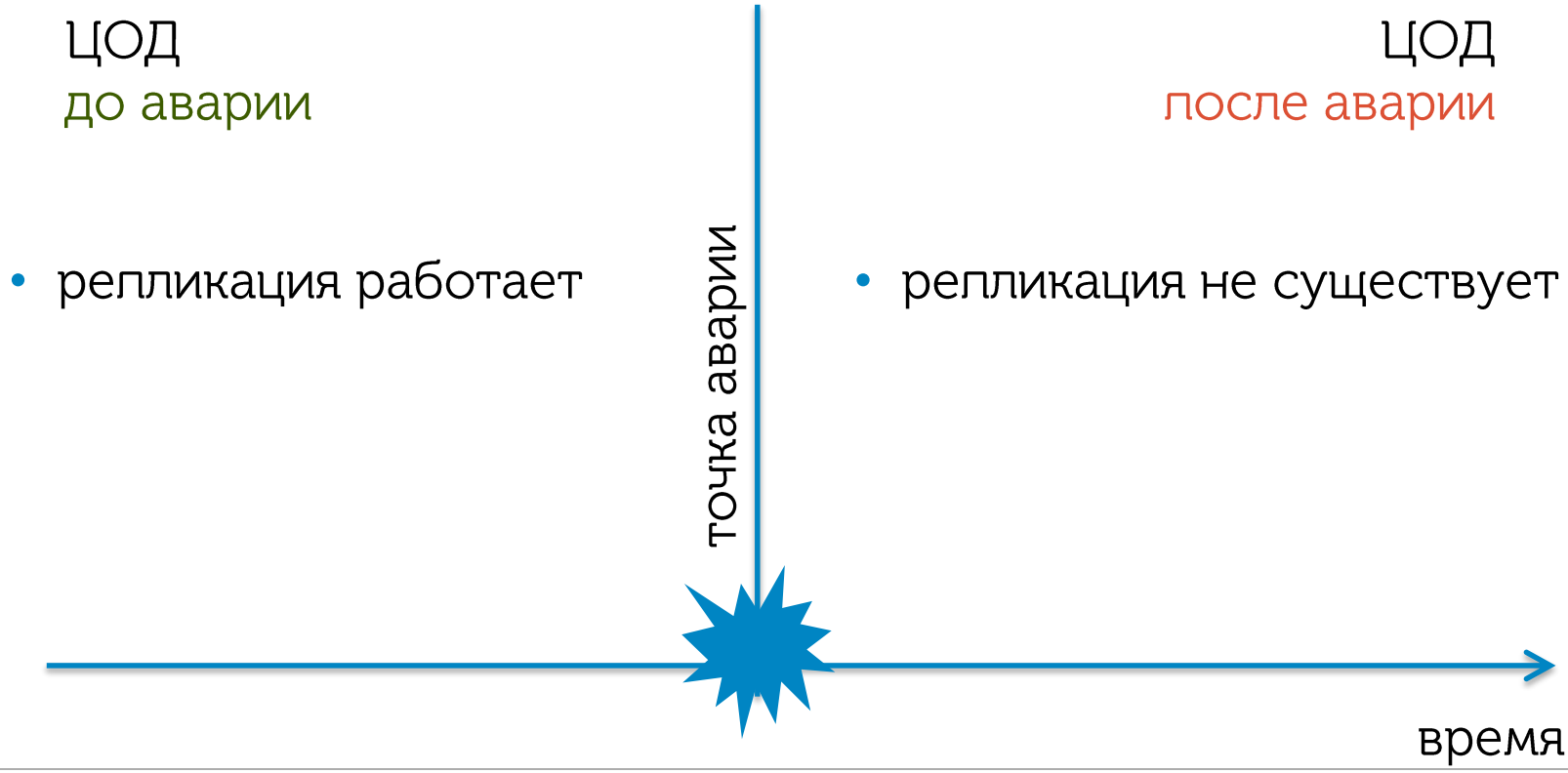
Очень важно осознать, что ЦОД до аварии и ЦОД после аварии – это два очень разных ЦОД.

В моменте после аварии метрокластера не существует.

Концептуальная основа
Disaster Avoidance vs Disaster Recovery



Концептуальная основа
Disaster Avoidance vs Disaster Recovery



Концептуальная основа
Disaster Avoidance vs Disaster Recovery

Концептуальная основа
Disaster Avoidance vs Disaster Recovery

Концептуальная основа
Disaster Avoidance vs Disaster Recovery

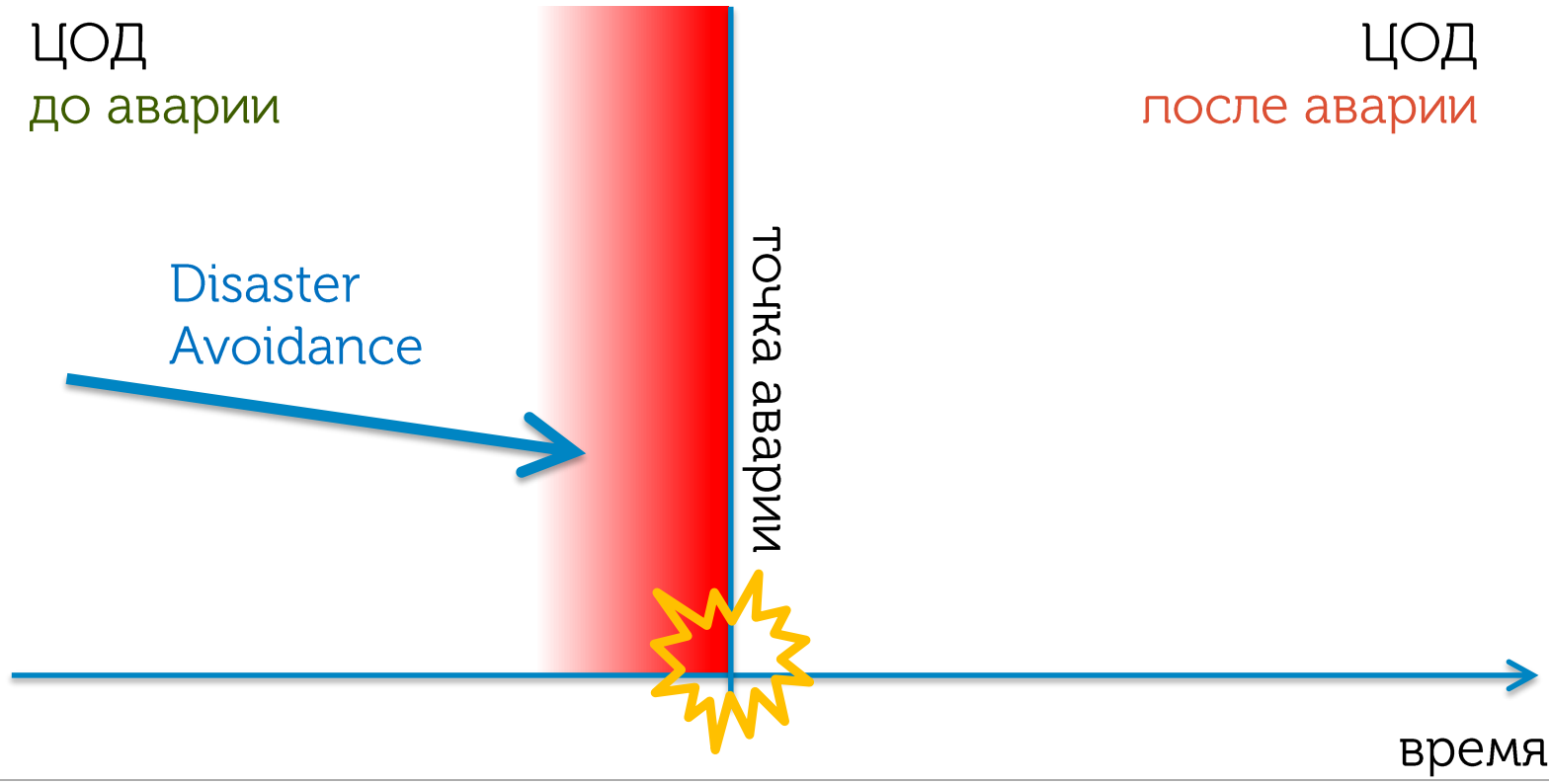
ЦОД
до аварии

- Балансировка нагрузки
- Успеть «убежать» = Disaster Aboidance

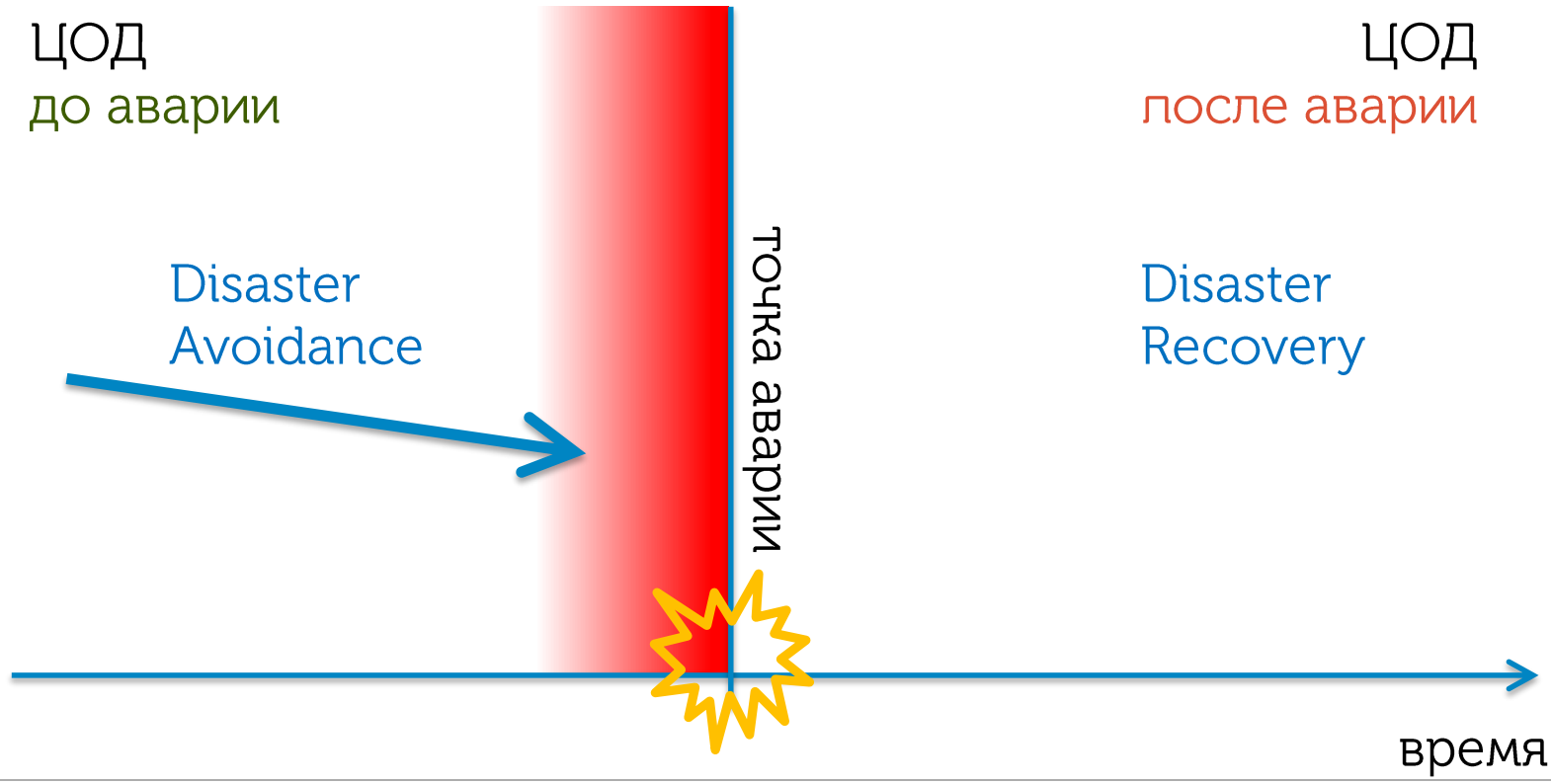
ЦОД
после аварии

- Восстановить работоспособность компании даже в худшем сценарии

Концептуальная основа
Disaster Avoidance vs Disaster Recovery



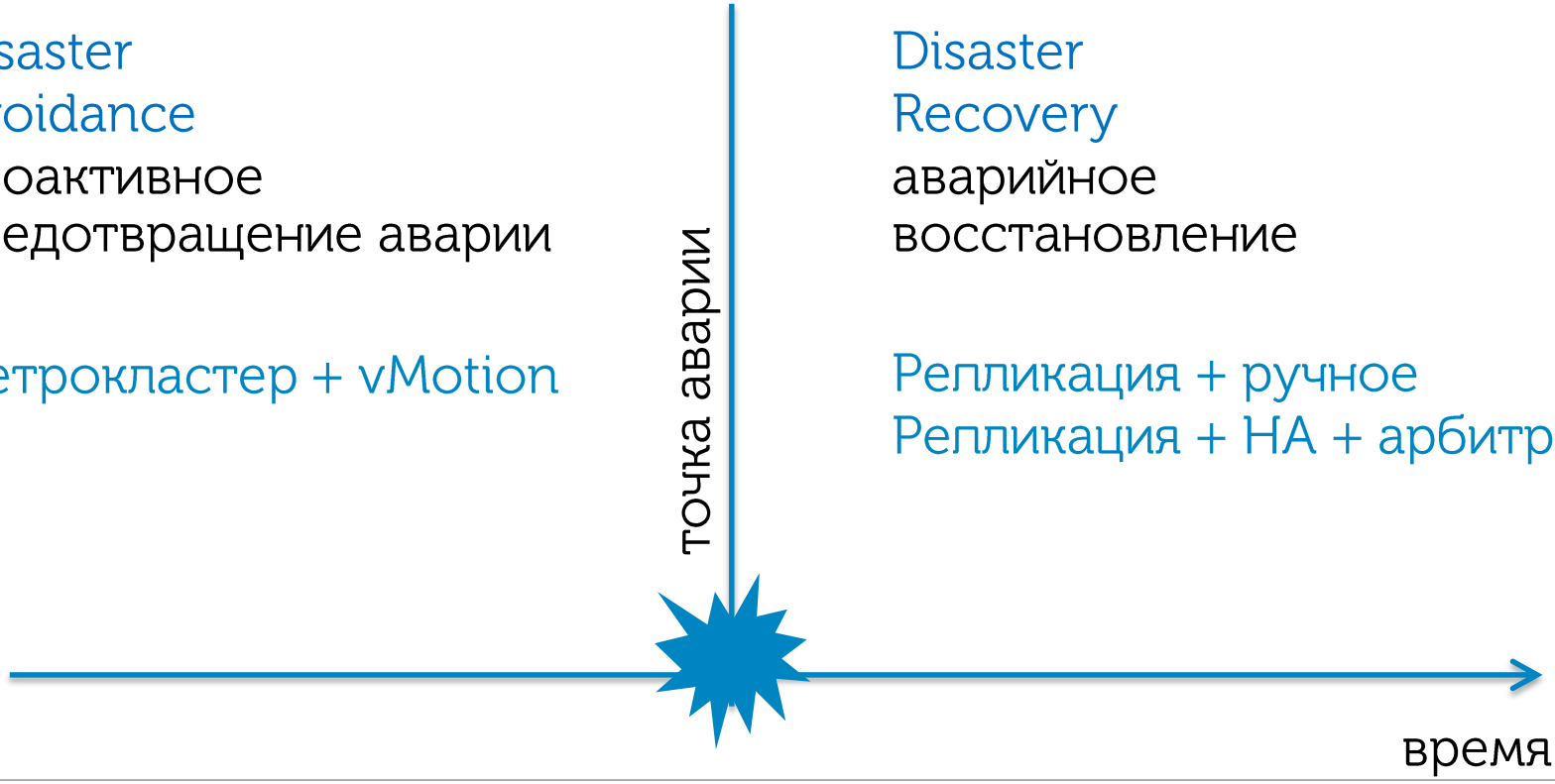
Концептуальная основа
Disaster Avoidance vs Disaster Recovery



Концептуальная основа
Disaster Avoidance vs Disaster Recovery

Disaster
Avoidance
проактивное
предотвращение аварии

Метрокластер + vMotion



Disaster
Recovery
аварийное
восстановление

Репликация + ручное
Репликация + HA + арбитр

Выводы:

1. Метрокластер – это хорошо, но и про DR-план не забываем

Выводы:

2. Метрокластер – это не замена репликации. Это и есть репликация. Просто с доп.возможностями – например, миграцией вирт.машин между площадками.

Выводы:

3. «Растянутой» площадки нет, если авария уже случилась. Поэтому серверы придется рестартовать, как и в старом добром DR-плане. И дело тут не в СХД. А в самом сервере приложений – он ведь тоже «упал»

Выводы:

4. Другой вопрос, что DR-план можно автоматизировать – дать серверам команду на рестарт автоматически. Если вы уверены, что это точно НЕ «сплит-брейн».

Выводы:

5. И автоматизировать рестарт можно при любой репликации – как «растянутой», так и «классической». Просто в растянутом варианте чуть меньше возни с ре-мапингом томов (так как тома уже были замапированы во времена «растянутого» состояния)

Выводы:

6. Кстати, «возня с ремапингом» не актуальна для заказчиков Dell Storage SC – в любом типе репликации. Потому что у них есть «Restore Points» и кнопка «Activate DR-план». То есть сценарий мапирования томов к серверам уже подготовлен заранее.

Техника.

Режимы репликации
Live Volume
Uniform \ Non-Uniform
AFO (Auto-Failover)

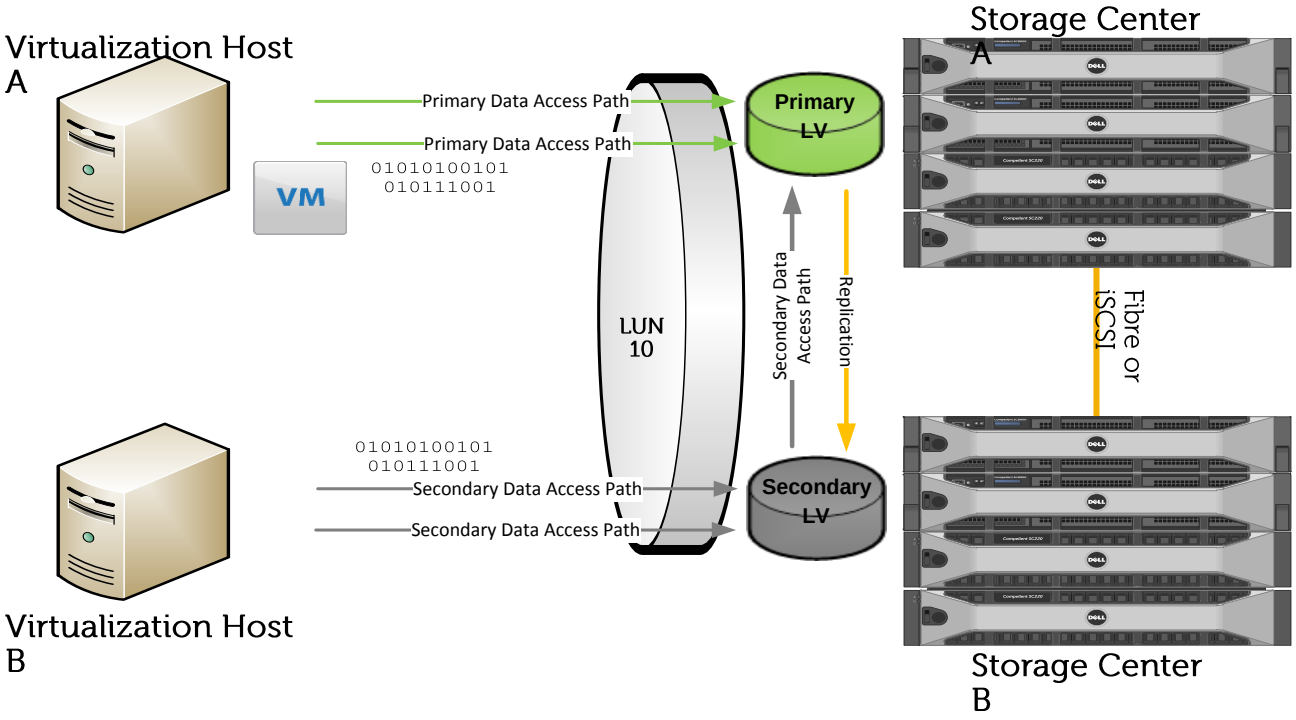
Режимы репликации в Dell Storage SC

1. Синхронная
2. Синхронная High Consistency
3. Асинхронная
4. Асинхронная по расписанию
5. Live Volume (sync \ async)
6. Live Volume AFO (только sync)
7. Live Volume LMR (трехсайтная с автослежением)

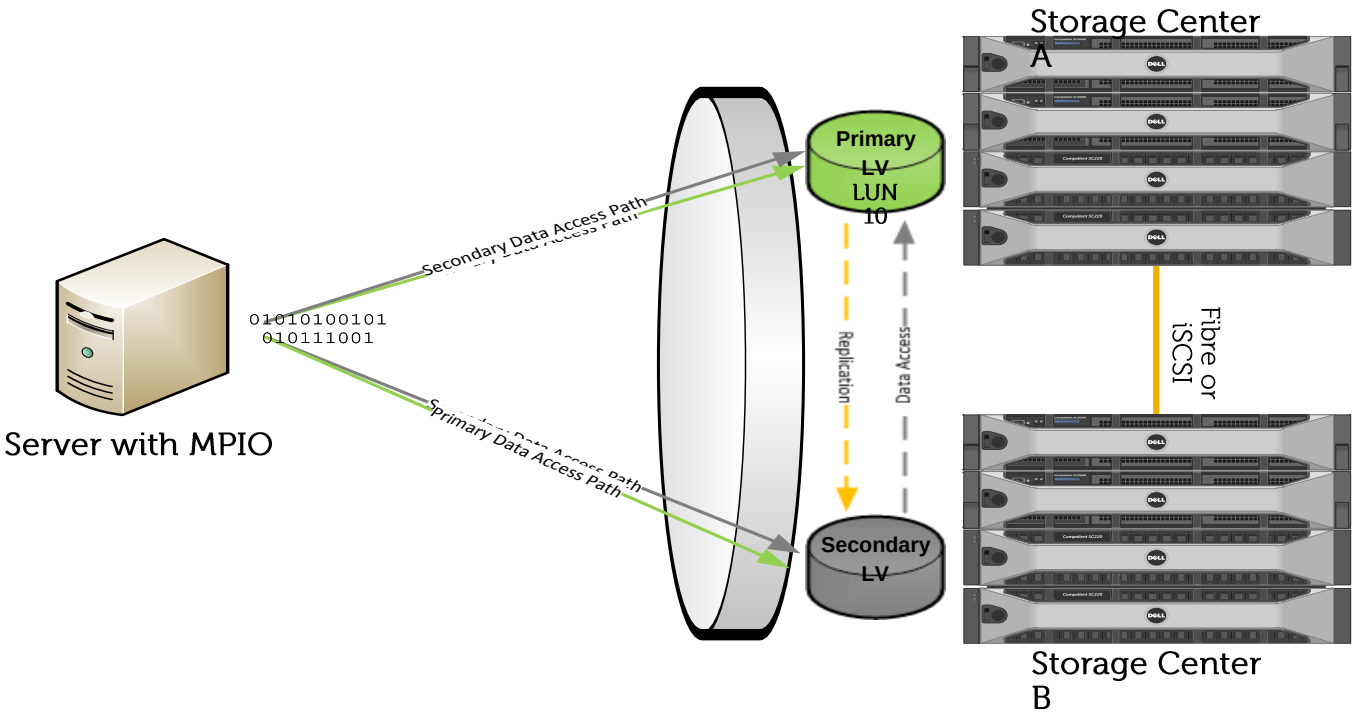
Режимы репликации в Dell Storage SC

	RPO	RTP	область применения
Синхронная High Availability	RPO = 0	RTO = ~ час	классика жанра
Синхронная High Consistency	RPO = 0	RTO = дни?	специфичные случаи
Асинхронная	RPO = ~мин	RTO = ~час	200+ км; IP-канал
Асинхронная по расписанию	RPO = от 5 мин до 24ч	RTP = ~час	1) snap как бэкап 2) дистрибуция данных
Live Volume	RPO = sync/async	RTO = sync/async	1) Телепортация ЦОД 2) Scale-Out
Live Volume AFO (Auto-Failover)	RPO = 0	RTO = 0 / мин	1) VMware Metro 2) Always-On Storage
Live Volume LMR (Livevolume-Managed Repl.)	RPO = 0 + async	RTO = AFO + async	трехсайтная репликация (sync + async)

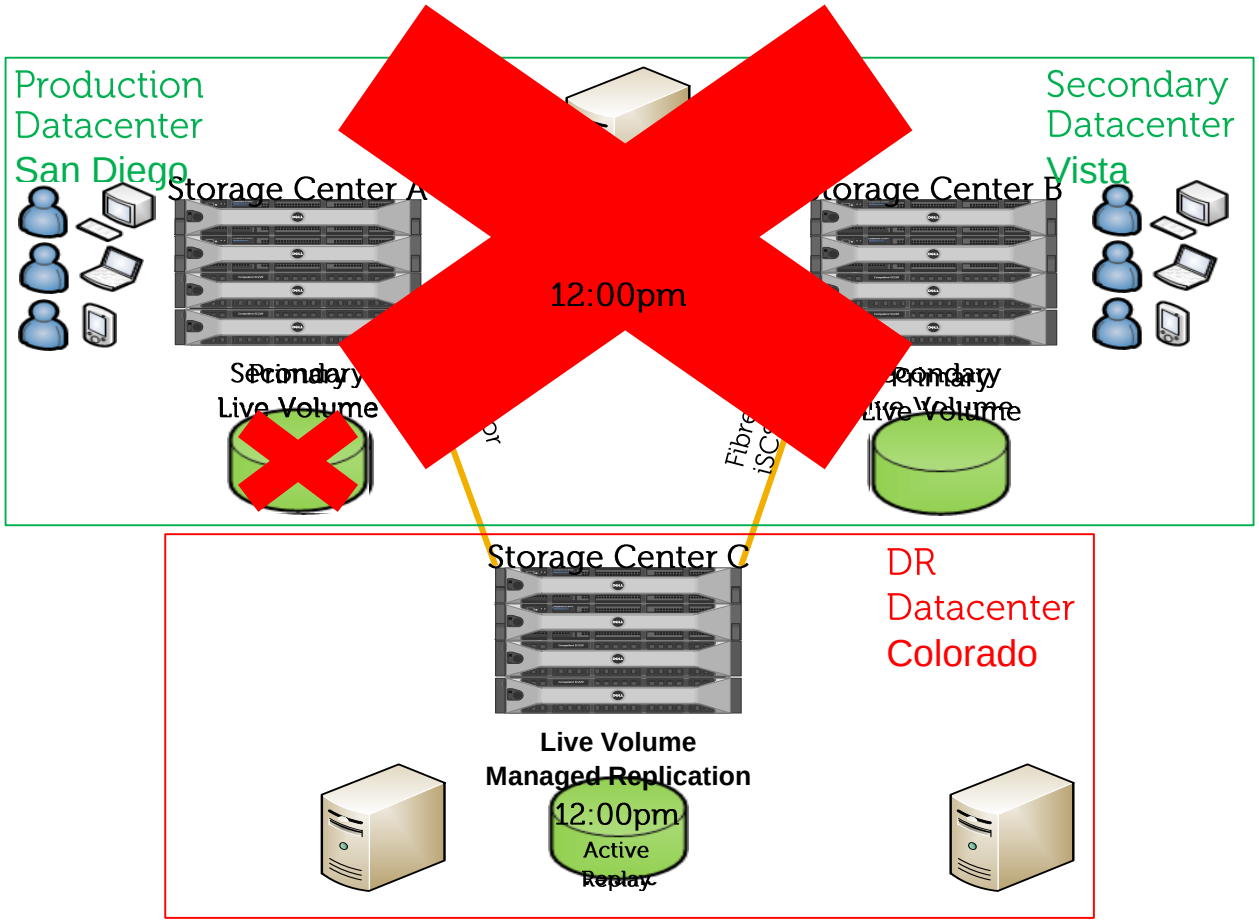
Live Volume


Virtualization Host A
VM
01010100101
010111001
Primary Data Access Path
Primary Data Access Path
LUN 10
Primary LV
Secondary LV
Secondary Data Access Path
Replication
Storage Center A
Fibre or iSCSI
Storage Center B
01010100101
010111001
Secondary Data Access Path
Secondary Data Access Path
Virtualization Host B

Миграция и «Always-On» СХД

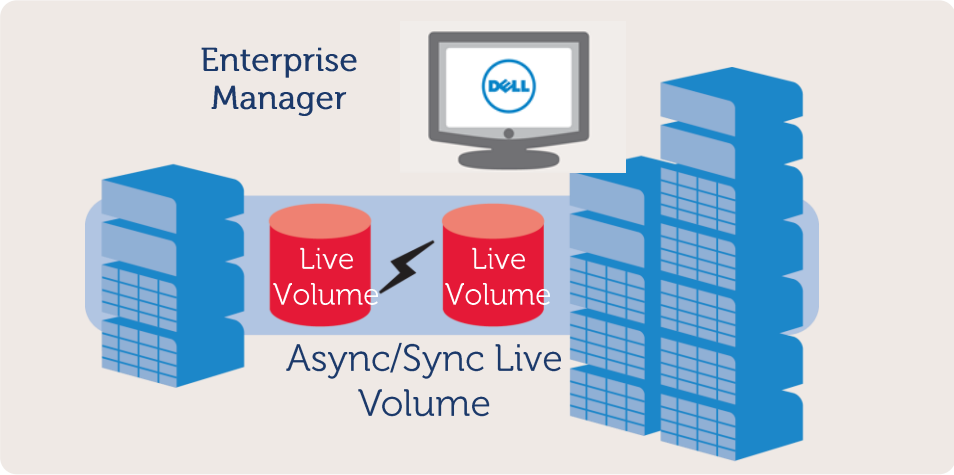


Трехсайтная репликация



Live Volume 6.7 – LV AFO (Auto Failover)

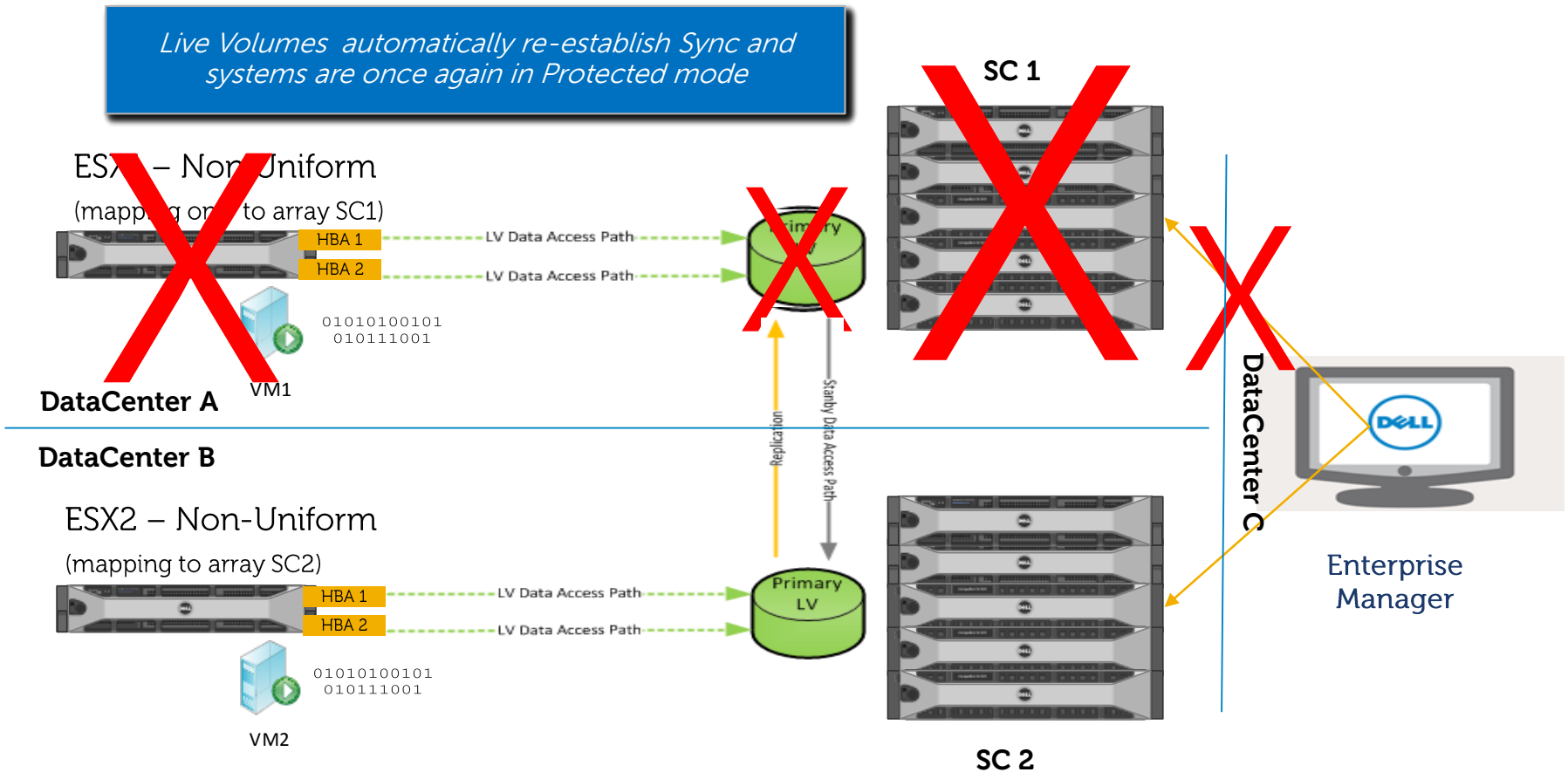
- Синхронная репликация
- Общий том, растянутый между двумя СХД
- Отсутствуют SPOF (единые точки отказа)
- Автоматическое переключение между СХД (в разных ЦОД) для непрерывного доступа к данным (и на чтение, и на запись, естественно)



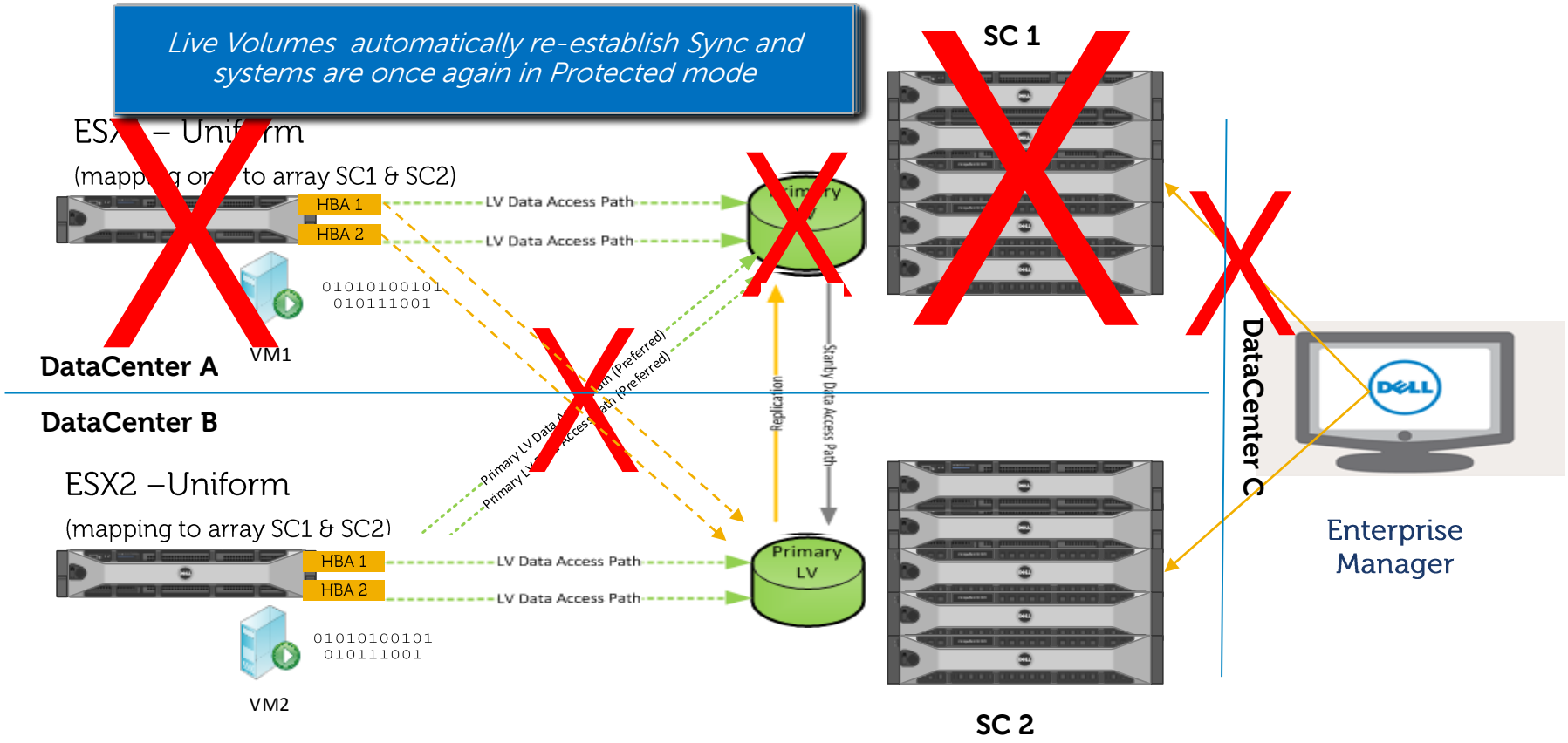
Live Volume AFO (Auto Failover)

High Availability	• 2 separate arrays each 99.999% uptime capable • Zero RPT/RTO
Quorum Witness	• Leverage Quorum Witness deployed in 3rd site • Witness is Dell Enterprise Manager • Requires Connectivity from each site
Requirement	• Low Latency/High Bandwidth Synchronous Replication Connectivity • Latency requirements still TBD, likely driven by application tolerances • Auto Failover between same Family of SC Series Controllers
vMSC Certification	• Solution Certified with VMware Metro Cluster • Initial release will be limited to VMware environments

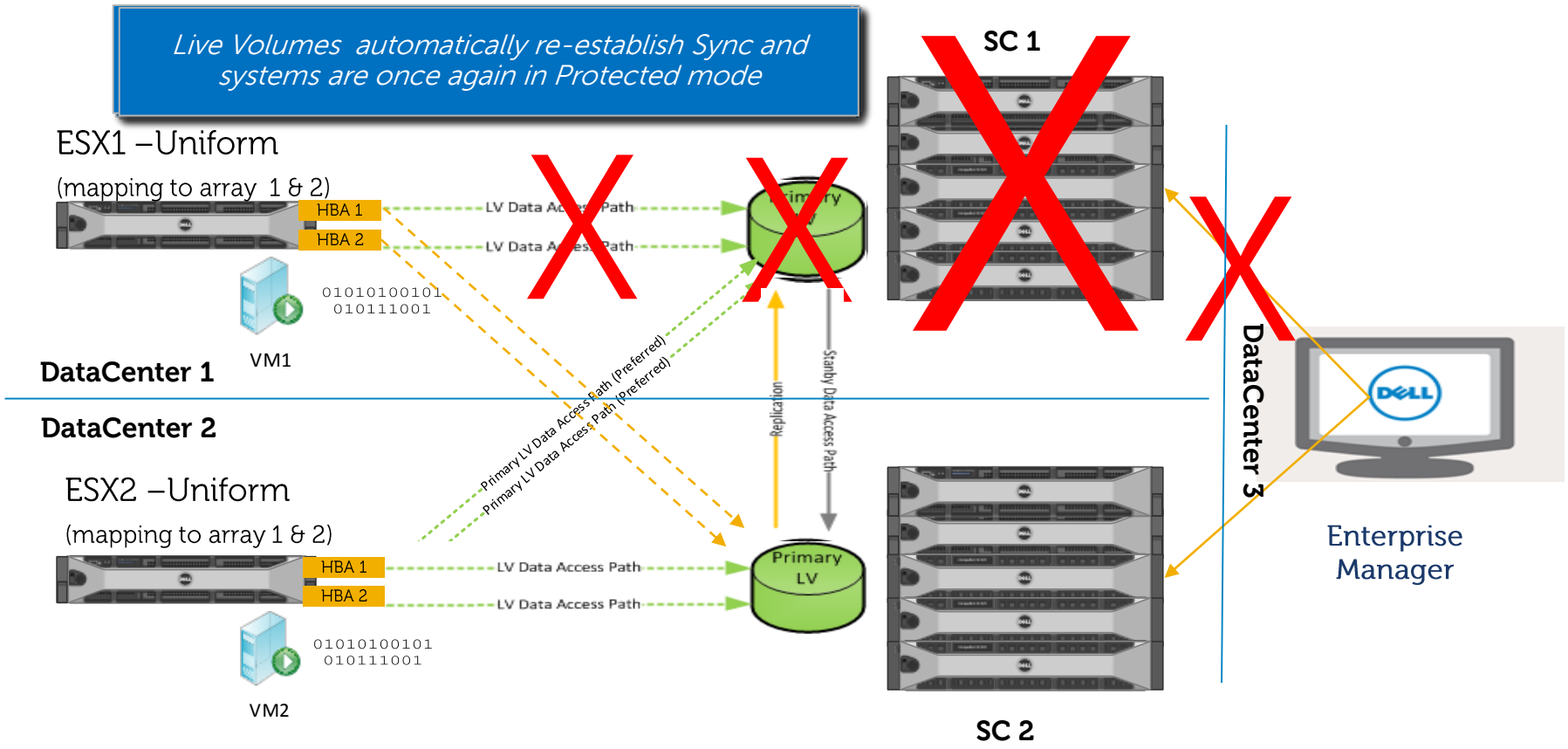
Live Volume Auto-Failover – Non-Uniform



Live Volume Auto-Failover - Uniform



Live Volume Auto-Failover – Uniform (Array Failure)



Uniform и Non-Uniform подключение

Non-Uniform:

+падение СХД = падение ЦОД. Соотв., серверы и приложения переезжают

+Зато не нужны кросс-линки

Uniform:

+при падении СХД приложения остаются работать, но нагрузка идет в РЦОД

Требования к Live Volume AFO

- SCOS 6.7
- Enterprise Manager 2015 R2
- Fibre Channel or iSCSI Replication
  - Synchronous High Availability Replication
  - Enough Bandwidth (Minimum of 250 Mbps)
  - Low Round Trip Latency on Replication link (Maximum of 10ms)
  - Connectivity to Tiebreaker 200ms of less of round trip latency
- vSphere 5.5/6.0
  - VMFS Datastores only, no physical mode RDMs (pRDMs)
  - HA configuration for PDL in 5.5/6.x and APD in 6.X (Live Volume best practices guide)
- Uniform or Non-Uniform Storage Presentation

Требования к Live Volume AFO

Hyper-V:

- поддержка начиная с SCOS 7.1
- в SCOS 6.7 нет поддержки AFO + Hyper-V
- в SCOS 6.7 есть поддержка только VMware ESX 5.5\6.0

И про кластер

Непрерывность работы MS-кластера (автопереключение на удаленную площадку) плюс автоподнятие в резервном ЦОД

+ это можно делать уже сейчас
+ MS Failover Cluster – и так уже имеет в себе кворум, и может запускать скрипты для активации Live Volume в случае падения ЦОД

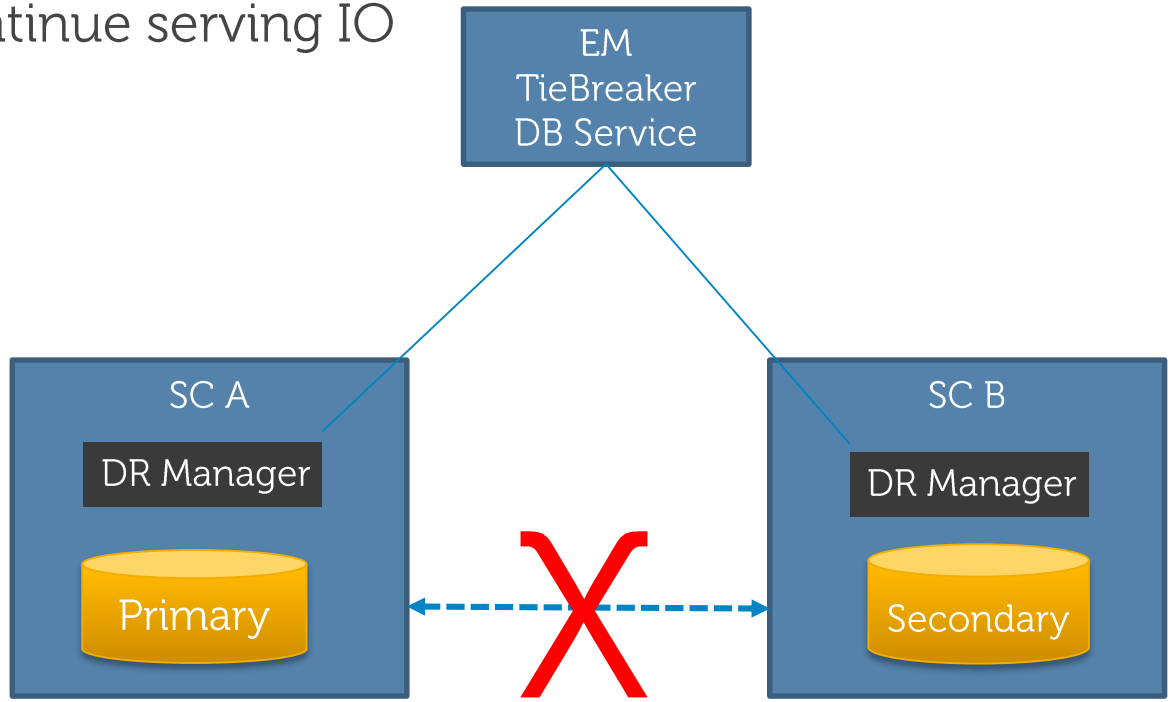
Сценарии отказоустойчивости

- Primary Array Fails
- Secondary Fails
- Network Partition between arrays
- Array issue taking Volume down (loss of multiple drives)
- EM Down
- EM Link down

Network Partition

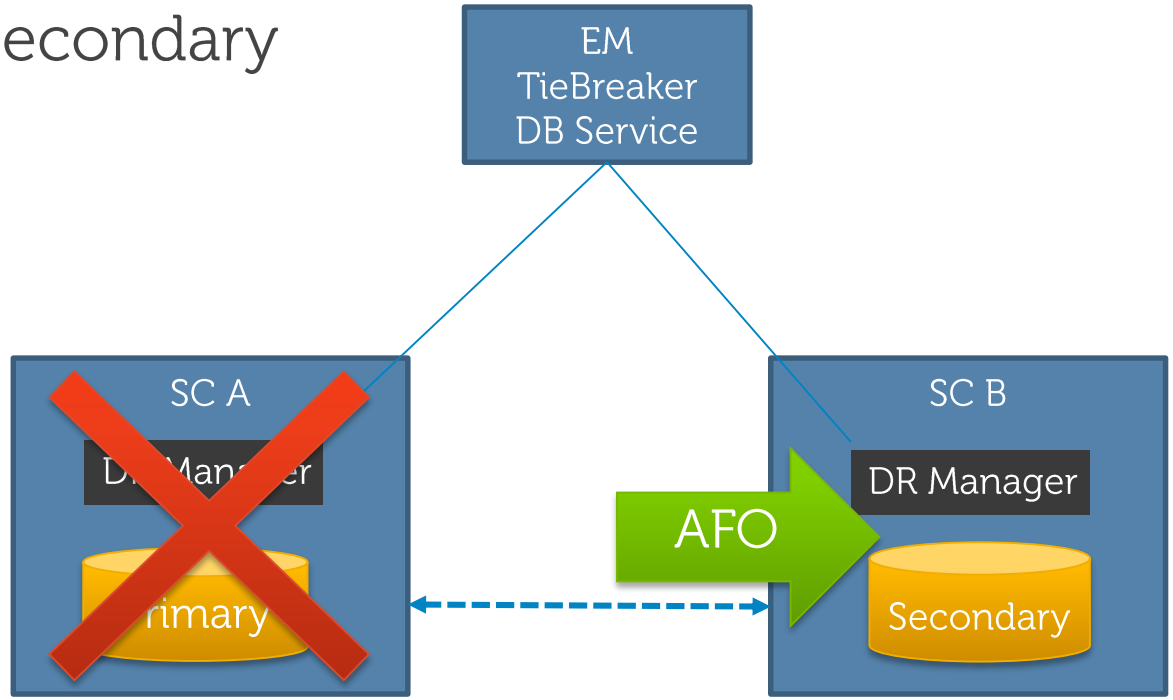
Replication/Live Volume Link Fails

- No Auto-Failover
- Primary Volumes continue serving IO



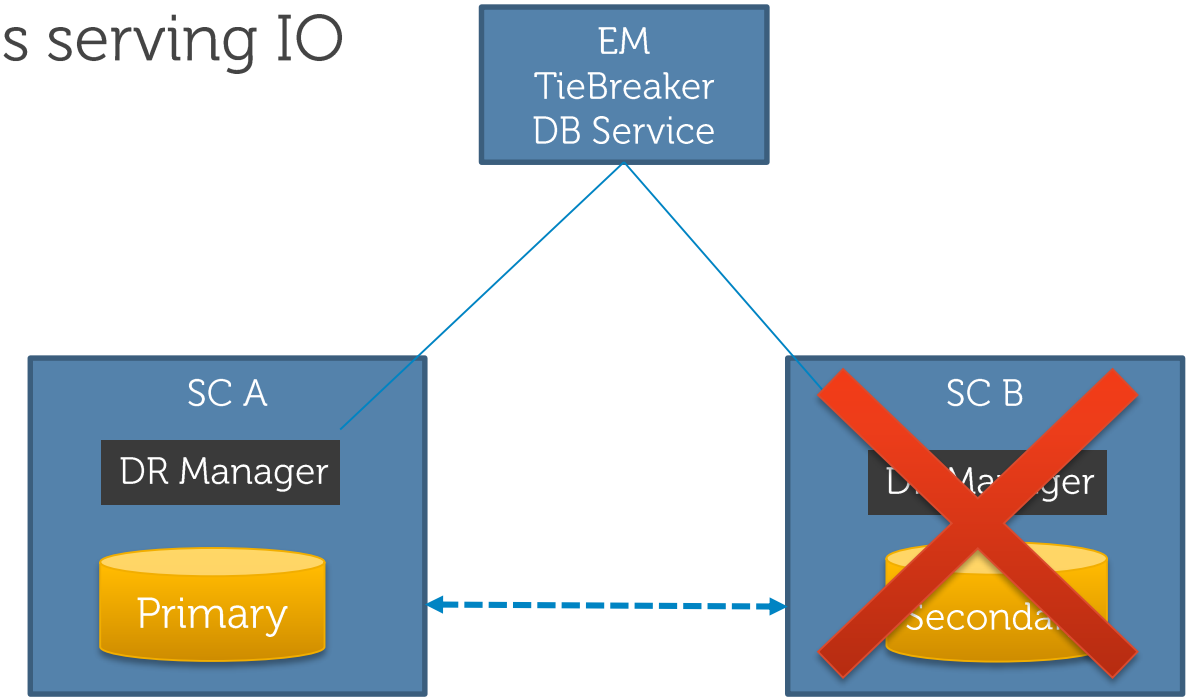
Array Failure Primary

- Primary Array Fails
- Auto-Failover to Secondary



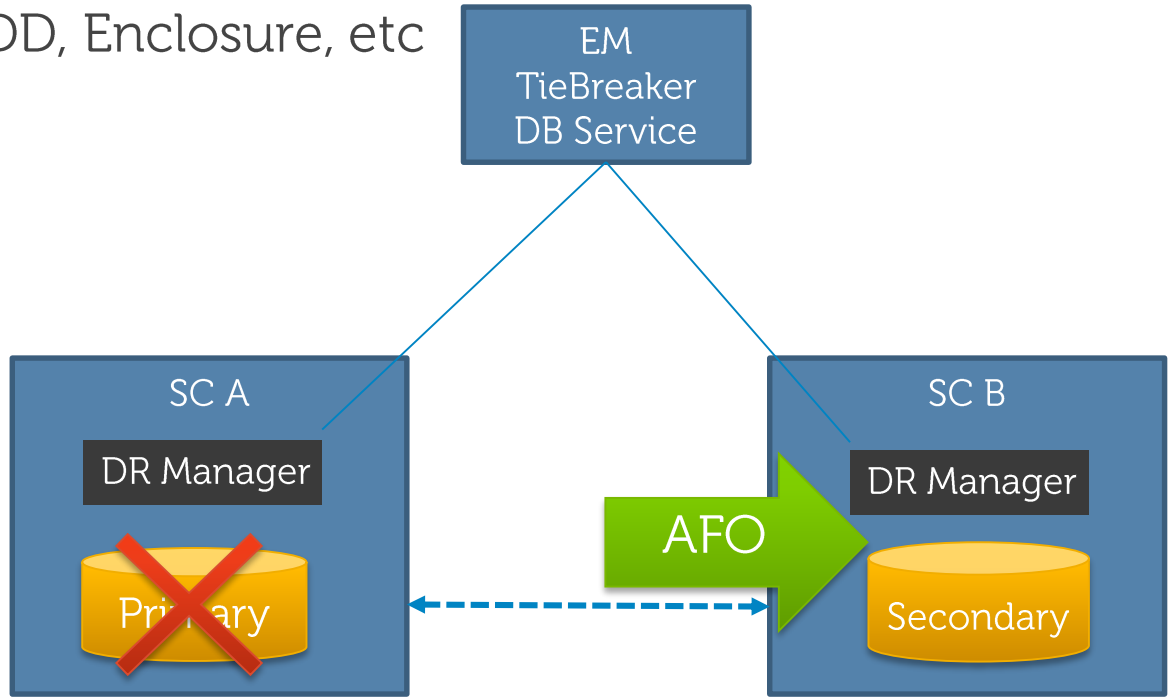
Array Failure - Secondary

- Secondary array fails
- Primary continues serving IO



Volume Outage

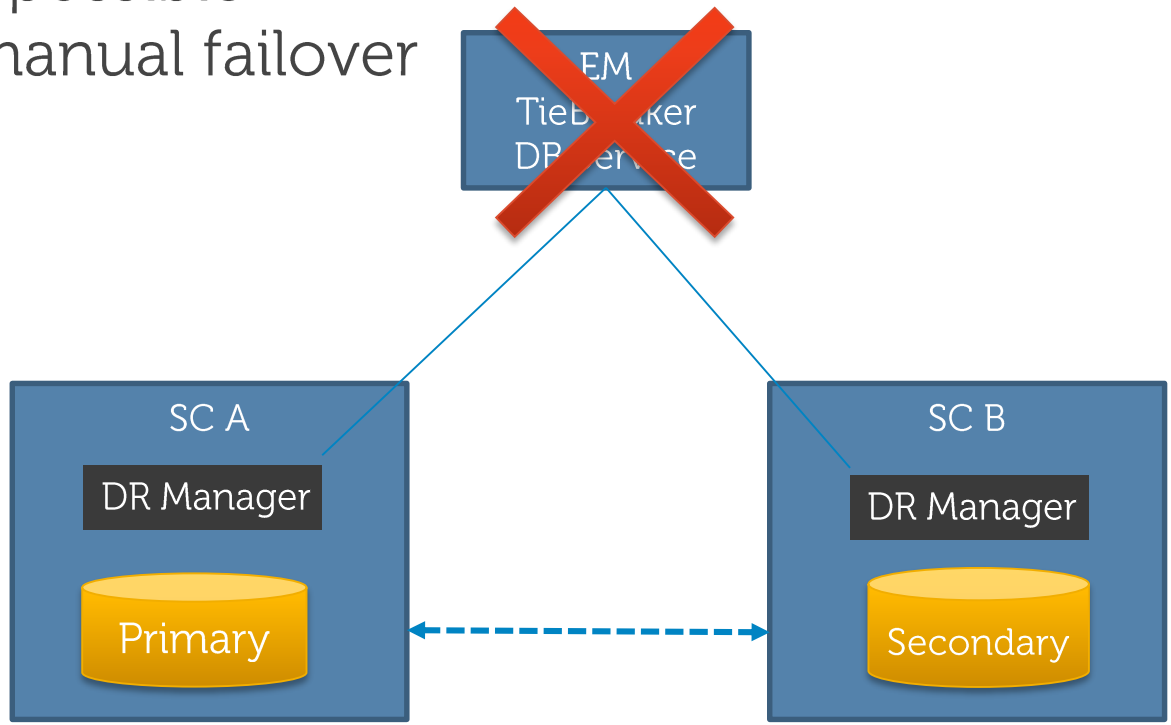
- Volume Outage
- Array is online but cannot serve IO for volume
  - Loss of multiple HDD, Enclosure, etc
- Auto-Failover



Enterprise Manager Failure

- EM Failure

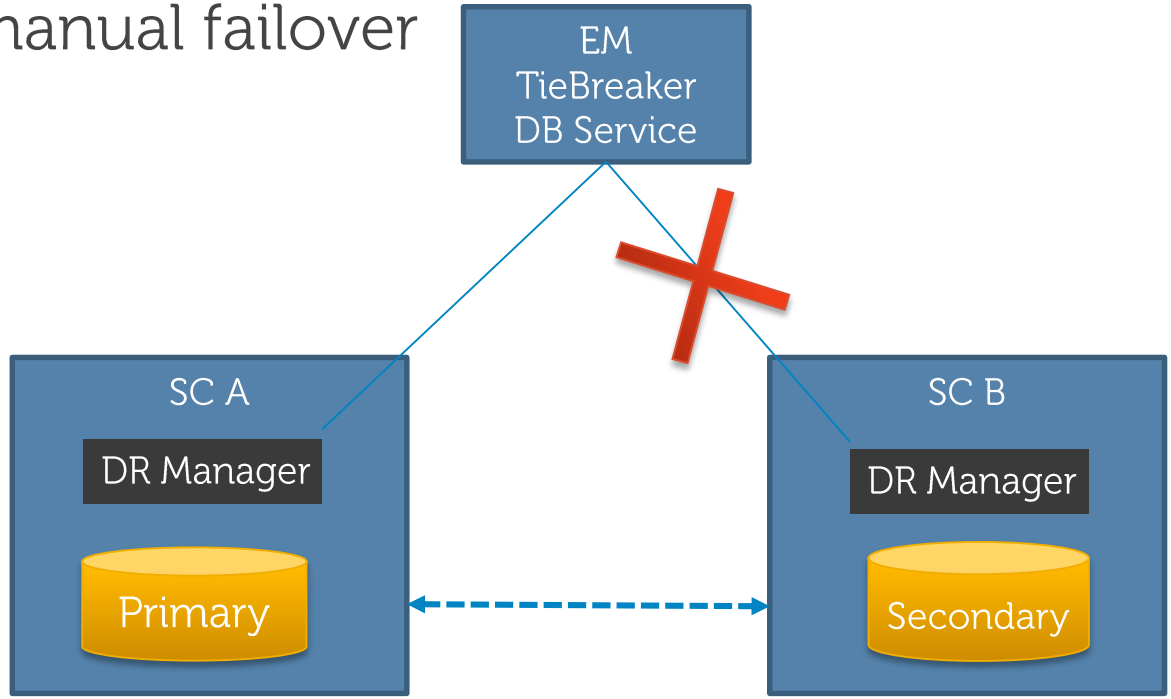
- Arrays continue operating normally
- Auto-Failover not possible
- Can still provide manual failover



EM Link Failure

- EM Link Failure

- Arrays continue operating normally
- Auto-Failover not possible
- Can still provide manual failover



Лицензирование:

- Все 7 типов репликации в одной лицензии
- Все 7 типов настраиваются в одном окне, и между режимами можно переключаться на ходу

- Сейчас:
  - SC9000 – «Вкл\Выкл» (1 лиц. на массив)
  - SC8000 – бандл «RIRA+LV». База (16 дисков) + экстеншены (с 17-го по 96-й диск с шагом 8)
  - SC4000 – бандл «Data Protection». База (48 дисков) + экстеншены (с 49-го по 96-й диск с шагом 24)
  - SCv2000 – доступна только асинхронная репликация, и только между SCv2000. «Вкл\Выкл» (1 лиц. на массив)

- В след. поколении SC:

  Единая схема лицензирования на всех моделях как сейчас у SC9000

Любопытные примеры

1. Live Volume как Scale-Out
   - пример из жизни

2. Live Volume для «доставания снэпшота из РЦОД»
   - пример из жизни

3. Live Volume для «Always-On» СХД с устойчивостью к двойным отказам
   - как идея на подумать…

Что мне нравится в Live Volume:

+ Очень, очень, очень просто настроить
+ Для заказчика это технически и экономически Risk-Free:
  - переключается между обычной репликой и растянутой – мышкой в настройках, в онлайне, без остановок
  - не требует перестройки архитектуры SAN
  - не требует сложных работ
  - не требует больших инвестиций
+ Поддерживает 100% функционала снэпшотов
+ Скриптуется PowerShell-ом